HOPE：面向电子学习和医疗保健的以人为中心的策略外评估

Ge Gao 北卡罗来纳州立大学 罗利，美国

Song Ju 北卡罗来纳州立大学 罗利，美国

Markel Sanz Ausin 北卡罗来纳州立大学 罗利，美国

Min Chi 北卡罗来纳州立大学 罗利，美国

arXiv:2302.09212v1 [cs.LG] 18 Feb 2023

摘要

强化学习（RL）已被广泛研究，用于提升电子学习和医疗保健等各种以人为中心的任务中的人类—环境交互。由于在这类任务中在线部署和评估策略风险很高，离线策略评估（OPE）对于制定有效策略至关重要。然而，在以人为中心的环境中，离线策略评估颇具挑战，因为潜在状态通常无法观测，只能观测到聚合奖励（例如学生的测试分数，或患者最终是否出院）。在本研究中，我们提出以人为中心的离策略评估（*HOPE*），以处理这类环境中的部分可观测性和聚合奖励。具体而言，我们在考虑部分可观测性的情况下，根据聚合奖励重建即时奖励，以估算期望总回报。我们为所提出的方法给出了理论界，并在真实世界的以人为中心任务中进行了广泛实验，包括脓毒症治疗和智能辅导系统。我们的方法能够可靠地预测不同策略的回报，并且无论采用标准验证方法还是以人为中心的显著性检验，表现都优于最先进的基准方法。

关键词

离策略评估、离线强化学习、以人为中心的环境、电子学习、医疗保健

1 引言

离策略评估（OPE）处于离线强化学习（RL）研究的核心位置 [11, 30, 48, 66],，其利用从历史数据中获得的先验知识来估计强化学习诱导的策略的表现。离策略评估在以人为中心的环境中尤为重要，因为执行不当的策略可能会造成高昂代价并带来危险。例如，在医疗保健领域，高估的策略可能无效，甚至因错误的治疗措施而增加死亡率。离策略评估要适用于以人为中心的环境，至少面临**两项主要挑战**。其一，以人为中心的环境的潜在状态通常是不可观测的 [8, 42],，从而形成部分可观测马尔可夫决策过程（POMDP）。具体而言，观测可能不包含足够的信息，无法完整重建潜在状态。例如，环境中的人群可能会随时间发生变化，其特征也可能因不同队列而异。在电子学习中，学生背景可能因学期而异；同样，在医疗保健领域，不同地区的医院中患者人口统计特征也可能发生变化 [33, 54]。**另一项挑战**在于，只有经过一段时间后才能观测到聚合奖励（或延迟奖励），而期间的所有即时奖励往往都缺失。电子学习和医疗保健中最恰当的奖励分别是学生学习表现和患者结局，但通常要等到整个轨迹完成后才能获得。这是因为学习和疾病进展都具有复杂性，难以逐时刻评估学生的学习情况或患者的健康状态。更重要的是，许多能够提升短期表现的教学或医疗干预措施，从长期来看可能并不有效。不同于经典的迷宫中的老鼠情境：智能体沿途获得微不足道的奖励，并在最终目标状态（食物）获得显著奖励；在电子学习和医疗保健中，沿途存在即时奖励，但这些奖励往往是不可观测的。

以往的离线策略评估工作 [3, 18, 30, 48]在 MuJoCo 等仿真环境中取得了出色的表现 [69]。然而，这些方法可能无法直接应用于以人为中心的环境，因为即时奖励缺失，且环境具有部分可观测性。一种可行的方法是假设奖励是稀疏的，仅用于指示某项任务是部分完成还是完全完成。稀疏奖励通常对应于某些特定任务的达成，例如机器人到达指定的路径点，而这些路径点几乎无法为即时步骤提供反馈 [58]。在许多以人为中心的任务中，交互结果可以通过逐步交互得到持续改善，每一步的即时奖励本身都可能具有实际意义。例如，在大学中，学生与智能导师互动期间，其认知结果和学习表现会逐步提高 [44]。研究发现，与稀疏奖励相比，即时奖励在评估决策结果方面更加有效 [5]。此外，在以人为中心的环境中，策略表现可能与环境的时域相关。例如，短期表现良好的医疗干预，长期来看可能并不有效 [5]。因此，要使离线策略评估在以人为中心的环境中有效运行，重构即时奖励至关重要。

另一方面，先前针对离线策略评估的评估指标通常是 Fu 等人 [14], Voloshin 等人 [74], 提出的误差指标（例如绝对误差、秩相关），而以人为中心的研究往往强调在实证研究中进行统计显著性检验 [22, 80]的必要性。例如，秩相关通过平均回报概括一组策略相对排名的表现。统计显著性检验可以说明我们所发现的关系在多大程度上仅仅是样本间的随机偶然所致，并且通常易于被领域专家

理解和采用 [22, 31]。用于验证以人为中心的环境中的离线策略评估的框架，可能需要超越误差指标。

在本研究中，我们提出了一种**以人为中心的离线策略评估（HOPE）**方法，以应对上述两个挑战。具体而言，该方法首先从聚合奖励中重构即时奖励。随后，使用重要性采样处理重构后的奖励，并估计期望总回报。任何现有的离线策略评估方法都可以与重构后的奖励联合使用，以估计期望回报。我们在医疗保健和电子学习领域的两个典型现实世界以人为中心的任务上验证了 HOPE，即例如脓毒症治疗和智能辅导系统（ITS），并通过针对以人为中心的环境的显著性检验扩展了现有的验证框架。总而言之，我们的工作至少有三项主要贡献：

- 据我们所知，HOPE 是首个同时应对部分可观测性和聚合奖励的离线策略评估方法。我们还为 HOPE 提供了理论界。
- HOPE 通过现实世界环境得到了广泛验证，这些环境包括电子学习和医疗保健领域的现实世界以人为中心的环境。结果表明，我们的方法优于当前最先进的离线策略评估方法。
- 我们在现有离线策略评估（OPE）验证框架的基础上引入显著性检验，以便全面研究和比较以人为中心的环境中的离线策略评估方法。

2 相关工作

在电子学习和医疗保健领域，研究人员已广泛探索利用历史用户交互数据 [20, 50]来学习强化学习（RL）策略。然而，在这些领域在线部署和评估策略风险很高，因为糟糕的策略可能会危及医疗保健领域人们的生命。因此，提出适用于以人为中心的环境的有效离线策略评估方法至关重要。离线策略评估用于评估给定历史数据时目标 策略的表现，这些历史数据取自（替代）行为 策略。期望累积奖励（ECR） [65], 等经典方法已应用于电子学习 [10]等现实场景。然而，ECR 在统计上并不一致，这对于高风险领域而言是一个重要问题 [42]。在实践中，研究人员发现，根据 ECR 选出的策略，甚至不如随机策略对真实学生有效（以人为中心的显著性检验） [10, 60]。

目前已提出多种 OPE 方法，可分为三大类 [74]: 1）逆倾向评分（IPS） [11, 56] 2） [12, 26, 30, 37, 40, 47, ; 2） 48, 53, 71, 75, 76] 3） [30, 66] 直接方法；3）混合方法。IPS 已在统计学 [27, 55] 和强化学习（RL） [56], 中得到广泛研究，其核心思想是利用 β 与 π 之间的重要性比率，对历史数据中的奖励进行重新加权。IPS 可以产生一致估计，并有多种变体，包括重要性采样（IS） [56]、加权重要性采样（WIS） [56],、逐步重要性采样（PDIS） [56], 和 PHWIS [11], 等。直接方法直接估计评估策略的价值函数。例如，FQE [37] 在功能上相当于批量Q学习的策略评估版本。DualDICE [48] 估计折扣平稳分布比率，该比率衡量 π 经历某个状态-动作对的可能性，并以该状态-动作对在离策略数据中出现的概率进行归一化。混合方法结合了 IPS 和直接方法的特点。例如，双重稳健（DR） [30] 是一种无偏估计量，它利用直接方法来降低 IS 产生的无偏估计值的方差。这些离线策略评估方法在仿真环境中取得了理想的性能。近期，一些研究提出了针对机器人抓取等现实世界任务的离线策略评估方法 [9, 19, 28, 51]。所提出的方法通常假设奖励可以从环境中观测到，或奖励是稀疏的。对于在现实世界以人为中心的环境中应用现有离线策略评估方法的有效性，相关研究仍然不足。

近期，研究人员认识到以人为中心的环境具有部分可观测性，并提出了针对部分可观测马尔可夫决策过程的离线策略评估方法 [2, 61, 64, 72]。他们假设潜在状态可能并不存在，并将其视为策略评估中的混杂因素。例如，Tennenholtz 等人提出了一种针对存在未观测混杂的部分可观测马尔可夫决策过程的方法，并在精心设计的合成环境中将其与重要性采样进行比较 [64]。结果表明，在一定程度的混杂条件下，他们提出的方法表现优于重要性采样。在实际应用中，例如我们针对真实学生开展的智能辅导实验，混杂因素未知且无法控制，同时还存在即时奖励缺失。我们的目标是利用真实世界电子学习和医疗保健环境中有限的可观测信息，在存在部分可观测性和即时奖励缺失的情况下，有效估计策略表现。

3 以人为中心的离线策略评估（HOPE）

问题定义： 我们将以人为中心的环境视为**部分可观测马尔可夫决策过程（POMDP）**，它是一个元组 $(\mathrm{S}, \mathrm{A}, \mathrm{O}, \mathrm{P}, \Omega\ R, \gamma)$。假设状态空间 S 未知。此外，A 表示动作空间，O是观测空间，P 定义了给定动作时从当前状态转移到下一状态的状态转移动态。观测模型 $p(o|s)\ \sim \Omega$ 也假设未知。 R是即时奖励函数，而 $r_t = R(s_t, a_t)$ 是即时奖励。 $\gamma \in (0, 1]$ 是折扣因子。每个回合都有有限时域 T。通常，在时间步 t，智能体从环境接收观测 $o_t \in \mathrm{O}$，然后遵循某个策略 π 选择动作 $a_t \in \mathrm{A}$: $\mathrm{O} \rightarrow \mathrm{A}$。随后，环境提供即时奖励 r_t 和下一次观测 o_{t+1}。在以人为中心的任务中，轨迹中大多数$t \in [1, T]$ 时间步的即时奖励 r_t 通常是未知的。不过，我们假设环境每隔 W 步会给出一个聚合奖励 $\mathring{r}_t = \sum_{i=t-W}^{t} \gamma^{i-t+W+1} r_i$，用于量化第 $t-W$ 步至第 t 步之间获得的即时奖励总和。

部分可观测马尔可夫决策过程中的离策略评估： 部分可观测马尔可夫决策过程中的离线策略评估旨在估计在目标 策略 π, $V^{\pi} = \mathbb{E}[\sum_{t=1}^{T} \gamma^{t-1} r_t | a_t \sim \pi]$下的期望总回报，所用数据是通过部署到环境中的行为 策略 $\beta \neq \pi$收集的历史数据。具体而言，历史数据 $\mathrm{D} = \{..., [..., (o_t, a_t, r_t, o_{t}) ...](i) ...\}_{i=1}^{N}$由一组 N轨迹组成，其中每条轨迹表示为 $\tau(i)$。

图1展示了 HOPE 如何应对以人为中心的环境中离线策略评估的两大挑战：部分可观测性和聚合奖励。具体而言，HOPE 首先根据聚合奖励推断初步的即时奖励（步骤1），然后 r_t 利用轨迹中的最近邻重构即时奖励

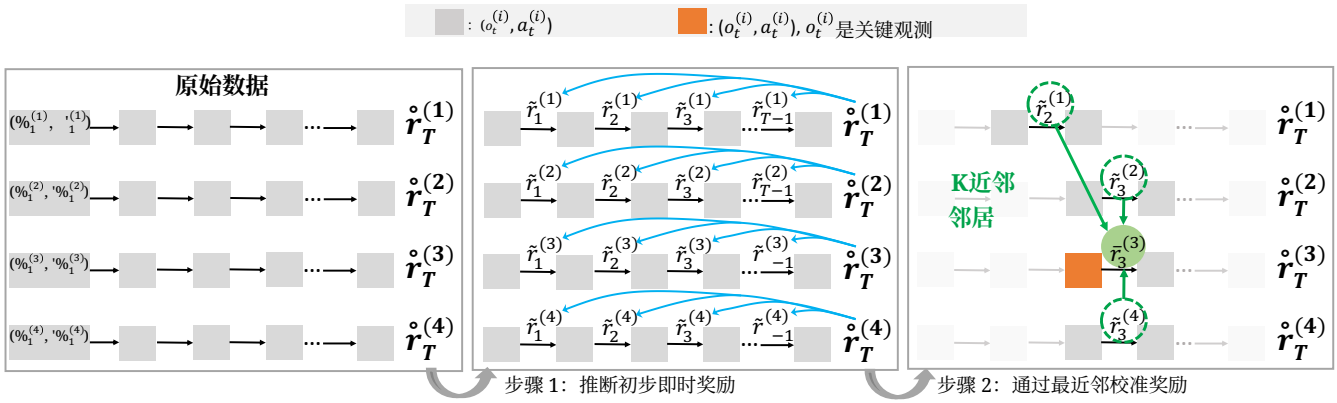


图1：即时奖励重构示意图。首先，根据聚合奖励推断初步奖励。然后，通过对最近邻的奖励取平均值，校准关键观测上的即时奖励。

D（步骤2），并将部分可观测性引入的不确定性纳入考虑。步骤2的思路源于先前研究 [17, 32, 78], 中的发现：具有相似潜在状态的人可能表现出相似的行为。随后，我们采用一种通用方法——加权重要性采样——来处理重构后的奖励并估计期望总回报V^{π}，因为这是最直接的方法，也有助于分离出我们提出的即时奖励重构框架所带来的改进。

3.1 通过最近邻重构即时奖励

在以人为中心的环境中，即时奖励通常不可用，而聚合奖励本身可能无法提供足够的信息，因而不能直接用于估计目标策略的期望总回报。尽管先前的研究已指出，估计即时奖励能够为训练强化学习策略提供有用信息 [5, 45],，但在离线策略评估的背景下，这一问题尚未得到广泛研究；这是因为以往的离线策略评估工作通常假设即时奖励可以从环境中观测到，或者即时奖励是稀疏的。在许多以人为中心的任务中，交互结果可以通过逐步改进而不断提升，每一步的即时奖励本身都能代表有意义的信息。因此，我们提出从历史数据中重构即时奖励，旨在丰富为离线策略评估提供的信息。

在我们所考虑的离线策略评估问题中，由于部分可观测马尔可夫决策过程，潜在状态无法被观测。此外，我们假设聚合奖励是在一个回合结束时获得的，遵循 $\mathring{r} = \sum_{t=1}^{T} \gamma^t r_t$，其中 T是环境的时域。在先前的工作中，可以使用函数逼近从聚合奖励 $\mathring{r}$ [5]推断即时奖励 r_t。因此，我们首先生成粗略的逐步初步重构奖励 $\tilde{r}$，以观测和动作作为输入，即 $\tilde{r} = f\ (o_t,\ a_t|\theta)$，并以即，

$$\min_{\theta} l(\theta) = \frac{1}{N}\sum_{i=1}^{N}(\mathring{r} - \sum_{t=1}^{T}\gamma^{t-1}\tilde{r})^2. \tag{1}$$

我们保持目标简单明了，并在 [4], 中采用标准训练方法，以便隔离表现提升的来源

。为简洁起见，在本文其余部分我们将 $\tilde{r}$ 称为初步即时奖励。在第4.5节所述的实验中，我们将其与把奖励视为稀疏奖励的方法进行比较，结果展示了初步即时奖励重构过程的有效性。

请注意，上文学习得到的映射 $f: \mathrm{O} \times \mathrm{A} \rightarrow \mathbb{R}$ 可能无法完美重构即时奖励函数 $R: \mathrm{S} \times \mathrm{A} \rightarrow \mathbb{R}$，因为观测模型 $p(o|s)$ 和状态空间 S 均被假设为未知。先前的研究发现，具有相似潜在状态的人类可能表现出相似的行为 [78,79], ，并且某些观测相比其他观测可能提供更关键的信息 [70]。因此，我们遵循这一直觉，提出一种补救方法：利用 $\tilde{r}_t$ 及其最近邻，从彼此相似的轨迹中重构即时奖励 r_t，以提高准确性。

具体而言，为捕获与潜在状态相关的信息，我们定义重构即时奖励 $\hat{r}_t$，同时考虑与相邻轨迹相关的平均奖励，以及相比其他观测可能提供更关键信息的观测；我们将这类观测称为关键观测，即：

$$\hat{r}_t = \mathbb{1}(o_t \in O^*)\bar{r}_t + \mathbb{1}(o_t \notin O^*)\tilde{r}_t, \tag{2}$$

其中， O^* 为关键观测集合，$\mathbb{1}(\cdot)$ 为指示函数。

我们首先介绍如何识别关键观测。给定 Q-函数 $Q^{\pi}(o,a)$，其表示在观测 o处采取动作 a的期望回报，任意两个动作之间的 Q^{π}差异（在相同的 o下）可用于量化最终结果差异的大小。因此，我们选择利用这种 Q差异来确定用于离线策略评估的关键观测。关键观测的正式定义如下。

定义 1（关键观测）。在离散动作部分可观测马尔可夫决策过程中，如果存在常数 h，使得 Q^{π}（o、 a） $a \in \mathcal{A}$的最大差异大于 h，即观测 o 是关键观测，

$$\max_{a',a'' \in \mathcal{A}} (Q^{\pi}(o,a') - Q^{\pi}(o,a'')) > h. \tag{3}$$

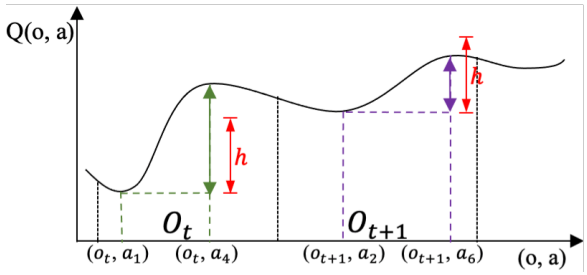


图2：根据 Q 值的最大差异识别关键观测。差异大于 Q 阈值 h 的 o_t 是关键观测。

在图 2 中，我们展示了如何使用阈值 h识别关键观测。第4.5节中的消融研究进一步证明了关键观测的必要性。

将轨迹 $\tau(i)$上时间步t 的重构即时奖励记为$\hat{r}(i)\,t$。我们将与关键观测 $o(i)\,t$ 相关联的平均奖励 $\bar{r}(i)\,t$ 定义为事件发生后邻近事件中出现的初步即时奖励 $\tilde{r}$ 的平均值，随后

$$\bar{r}_t^{(i)} = \frac{1}{K}\sum_k \tilde{r}_{t'}^{(k)}, \tag{4}$$

其中，求和是在所有 $\mathring{r}(k)\,t'$ 上进行的， K 是算法1对应的最近邻总数。具体而言，如果两条轨迹在观测空间和动作空间上的访问分布相似，则定义它们为邻居，随后

$$d(\tau^{(i)}, \tau^{(k)}) = similarity(o^{(i)}, o^{(k)}) + similarity(a^{(i)}, a^{(k)}) \tag{5}$$

其中，可以通过某些度量（例如Kullback-Leibler（KL）散度 [36]）计算分布的相似度。然后，如果两个观测分别来自两条相邻轨迹，并且共享最相似的观测，则定义这两个观测为邻居，因为从相似轨迹中共享相似观测的时刻可能代表人类潜在状态 [78, 79]。为了计算两个观测之间的相似度，可以使用欧几里得距离等距离度量。通过适当处理平局（例如选择最早的观测），我们总能在给定的相邻轨迹上找到唯一一个邻近观测。算法1总结了如何寻找 $o(i)\,t$ 的K个最近邻。在我们的实验中，第4.5节将所提出的最近邻距离与随机选择进行了比较，结果表明，所提出距离的表现更优。

3.2 HOPE

基于针对以人为中心的环境中相关挑战所提出的解决方案，我们通过将估计的即时奖励与加权重要性采样（WIS）相结合，提出了以人为中心的离线策略评估（*HOPE*）。令 $w_i := \Pi_{t=1}^{T}\pi(a(i)\,t\ |o(i)\,t)/\beta(a(i)\,t\ |o(i)\,t)$ 表示一个重要性权重，即在目标策略下前 T 步 $\tau(i)$ 发生的概率 π除以其在行为策略下的概率 β [56]。HOPE 对于一个

算法1： 查找 o（i）t 的K近邻。

输入： 历史数据D，超参数 K。

开始：

1：初始化空集Z_t^i。

2：**对于**D中的每条轨迹 τ（j）$\in$ **执行**

3：　按照（5）计算距离 $d(\tau$（i）τ（j）$)$。

4：**结束循环**

5：将 $d(\tau$（i）τ（j）$)$ $j \in N$按非递减顺序排列，取出前 K 个元素，并找到相应的轨迹，作为 τ（i）的相邻轨迹。

6：**对于** τ（i）**的**每条相邻轨迹 τ（k），执行

7：　**对于**每个观测 $o(k)_{t'}$ $\in \tau(k)$ **执行**

8：　　计算距离 $l^{ik}_{tt'}$ 与 $o(k)_{t'}$ 和 $o(i)_t$。

9：　**结束循环**

10：　按非递减顺序排列 $l^{ik}_{tt'}, t' \in T$，取第一个元素并找到对应的观测 $o(k)_{t'}$ 作为 $o(i)\,t$ 的邻近观测。

11：　$Z_t^i \leftarrow Z_t^i \cup o(k)\,t'$

12：**结束循环**

13：返回集合 Z_t^i。

目标策略 π 为

$$V^{\pi}(\text{HOPE}) = \frac{1}{\sum_{i=1}^{N} w_i}\sum_{i=1}^{N} w_i \sum_{t=1}^{T} \gamma^{t-1}\hat{r}_t^{(i)}, \tag{6}$$

我们选择在此使用 WIS，因为这是最直接的方法。此外，这有助于隔离我们提出的即时奖励重构框架所带来的改进来源。先前的研究指出，与实践中的某些无偏离线策略评估方法（例如重要性采样（IS））相比，WIS 较低的方差可能会使期望平方误差获得更大幅度的降低，而这种降低超过了其偏差所带来的额外误差 [68]。我们在第 4 节中的真实世界实验结果进一步支持了这一点。

一般而言，重要性权重假设评估策略 π 的支持集是行为策略 β 支持集的子集，而假设 1 [67] 强制满足了这一条件：

假设1。 如果 $\pi(a|o) \neq 0$，那么 $\beta(a|o) \neq 0$，其中 $a \in \mathcal{A}$ $o \in O$。

HOPE 的上下界。 我们将 δ^{ik} 定义为以下事件： $\tau(k)$ 是 $\tau(i)$ 的一个邻居。此外，我们将邻居事件的计数定义为 $K = \sum_{k=1}^{N}\mathbb{1}(\delta^{ik})$，其中 $\mathbb{1}(\cdot)$ 是指示函数。随后，我们构造最近邻矩阵： $M^{ik} = {}^{\mathbb{1}}(\delta^{ik})\,K$。由于 $N \times N$矩阵 M^{ik}可以根据数据计算得到，并可用于使用以下命题计算所有观测-动作对的估计即时奖励。

命题1。 对于数据中的所有转移，相应观测-动作对的估计即时奖励为

$$\hat{r}_t^{(i)} = \sum_k M^{ik} \equiv [\boldsymbol{Mu}]_i \tag{7}$$

证明。在时间戳 t 的轨迹 i 上给定的奖励，即对满足 $\delta_{tt'}^{ij}$ 的时间戳 t' 处轨迹 j 上的所有奖励取平均，可写为 $\frac{1}{K}\sum_{(j,t')}$: $\delta_{tt}^{ij\prime} = \sum_{j,\ t'} \mathbb{1}(\delta_{tt}^{ij\prime})_K = \sum_{j,\ t'} M_{tt'}^{ij}$。因此，假设 $u(o,a)$ 是定义在观测-动作空间上的函数，并且 u 是一个向量，其中包含对于每个 $(o(i)\ a(i))$ 的数量 $u_i = u(o(i)a(i))$，则 $u(o(i)\ a(i))$ 的最近邻估计为 $[\mathrm{Mu}]_i$。□

在 $\tilde{r} \in [\tilde{r}_{lb}, \tilde{r}_{ub}]$ 的情况下，根据命题 1 得到 $\hat{r} \in [\tilde{r}_{lb}, \tilde{r}_{ub}]$。将轨迹 $\tau(i)$ 的回报记为 $G(\tau(i))$。按照 [67]，我们还将 $G(\tau(i)) \in [0,1]$ s.t. $G(\tau(i)) := \frac{(\sum_{t=1}^{T}\gamma^{t-1}\hat{r}_t^{(i)}) - G_{lb}}{G_{ub}-G_{lb}}$，以表示对轨迹 $\tau(i)$ 优劣程度的量化。于是，HOPE估计可写为 $\frac{1}{\sum_{n=1}^{N} w_i}\sum_{i=1}^{N} w_i G(\tau(i)) = \frac{\sum_{i=1}^{N} w_i G(\tau^{(i)})}{\sum_{i=1}^{N} w_i}$。随后，HOPE估计的上界和下界可利用以下引理进行计算，其证明见 [67]。

引理1。 设 π 和 β 为满足假设*1*的任意策略，则对于任意常整数 $m \geq 1$，$\mathbb{E}[\prod_{t=1}^{m} \frac{\pi(a_t|o_t)}{\beta(a_t|o_t)} | \tau \sim \beta] = 1$.

根据引理1，随着样本数量增加，HOPE的分母趋近于 n。HOPE估计量被限制在 $[0,1]$, 以内，因此对于所有 π 和 β，$\rho_{lb}^{HOPE}(\pi,\ \beta) = 0$ 以及 $\rho_{ub}^{HOPE}(\pi,\ \beta) = 1$，其中 $\rho(\pi) = \mathbb{E}[G(\tau)|\tau \sim \beta]$。

*HOPE*的一致性。 我们还表明，如果存在单个行为策略（定理*1*），或者存在多个满足技术要求的行为策略（定理*2*），则HOPE是 $\rho(\pi)$的一致估计量，这延续了 [67]的研究工作。

定理1。 如果假设*1*成立且只有一个行为策略，那么*HOPE*是 ρ（π）的一致估计量。

证明。当只有一个行为策略时，通过将分子和分母同时乘以 $\frac{1}{N}$，HOPE的估计可以改写为 $\frac{\frac{1}{N}\sum_{i=1}^{N} w_i G(\tau^{(i)})}{\frac{1}{N}\sum_{i=1}^{N} w_i}$。此时分子等于$IS^{\pi}$，正如先前工作 [56, 67]所证明的，这是 $\rho(\pi)$的一致估计量，因此分子几乎处处收敛于 $\rho(\pi)$。对于分母，根据引理1，我们有

$$\mathbb{E}[\prod_{t=1}^{T}\frac{\pi(a_t^{(i)}|o_t^{(i)})}{\beta(a_t^{(i)}|o_t^{(i)})}] = 1, \text{for all } i \in [1,\ldots,N]. \tag{8}$$

由于只有一个行为策略，对于每个 $\prod_{t=1}^{T}\frac{\pi(a_t^{(i)}|o_t^{(i)})}{\beta(a_t^{(i)}|o_t^{(i)})}$ $i \in (\ )$，$[1,\ldots,N]$ 每一项均同分布。根据Khintchine强大数定律 [59], ，我们有

$$\frac{1}{N}\sum_{i=1}^{N}\prod_{t=1}^{T}\frac{\pi(a_t^{(i)}|o_t^{(i)})}{\beta(a_t^{(i)}|o_t^{(i)})} \xrightarrow{\text{a.s.}} 1. \tag{9}$$

根据几乎必然收敛的性质， [29], HOPE几乎必然收敛到 $\rho(\pi)$, 因此HOPE是$\rho(\pi)$的一致估计量。□

我们还提供了存在多个行为策略时 HOPE 一致性的证明。

定理2。 如果假设*1*成立，并且存在常数 $\epsilon > 0$ 使得 $\beta_i(a|o) \geq \epsilon$对所有 $i \in \{1、\ldots、N\}$ 以及 $(a|o)$ 均成立，其中$\pi(a|o) \neq 0$, 那么当存在多个行为策略时，*HOPE* 是 ρ（π）的一致估计量。

证明。当存在多个行为策略时，与定理1的证明类似，分子等于 IS^{π}。分子几乎必然收敛到 $\rho(\pi)$。对于分母，根据引理1可知，等式8成立。每一项 $\prod_{t=1}^{T}\frac{\pi(a_t^{(i)}|o_t^{(i)})}{\beta_i(a_t^{(i)}|o_t^{(i)})} \in [0, \frac{1}{\epsilon^T}]$ 的方差均有界。因此，根据Kolmogorov强大数定律 [59], ，等式9几乎必然成立。因此，当存在多个行为策略时，HOPE是 $\rho(\pi)$的一致估计量。□

4 实验

我们在两个真实世界的以人为中心的任务——脓毒症治疗和智能辅导——上开展实验，以验证我们提出的方法。由于我们关注的是以人为中心的环境，因此不对控制任务（如D4RL）开展实验 [13]。

4.1 基准方法

我们采用九种 最先进的基准方法，涵盖了离线策略评估（OPE）中已探索的多种方法：四种逆倾向评分方法，包括重要性采样（IS） [56], 加权重要性采样（WIS） [56], 逐步重要性采样（PDIS） [56], 和逐步时域加权重要性采样（PHWIS） [11]。对于 PHWIS，我们采用 [11], 中的 PHWIS-Behavior，因为我们假设轨迹长度不取决于用于生成轨迹的策略。两种直接方法：拟合Q评估（FQE） [37] 和双重平稳分布校正估计（DualDICE） [48]。对于 FQE，我们按照 [37], 的方法，通过从 $Q(o,a)$进行自助法，训练一个神经网络来估计评估策略 π_e的价值。对于 DualDICE，我们使用其原始论文中公开的代码。三种混合方法，包括双重稳健（DR）、加权双重稳健（WDR） [66],和 MAGIC。 [66], MAGIC 会在一组线性组合中寻找最优组合，这些组合通过改变 WDR 与直接方法之间的切换点而形成。对于 MAGIC，我们使用 [74] 的实现。

4.2 验证离线策略评估表现

在本研究中，我们使用两类程序来验证离线策略评估方法的性能。我们采用先前离线策略评估方法中常用的标准指标，包括绝对误差、regret@1 和斯皮尔曼秩相关系数 [63]

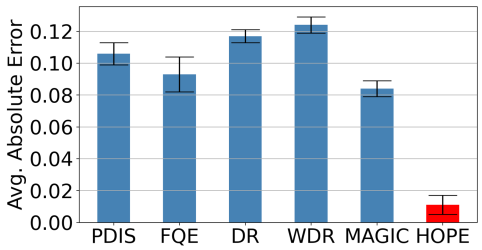
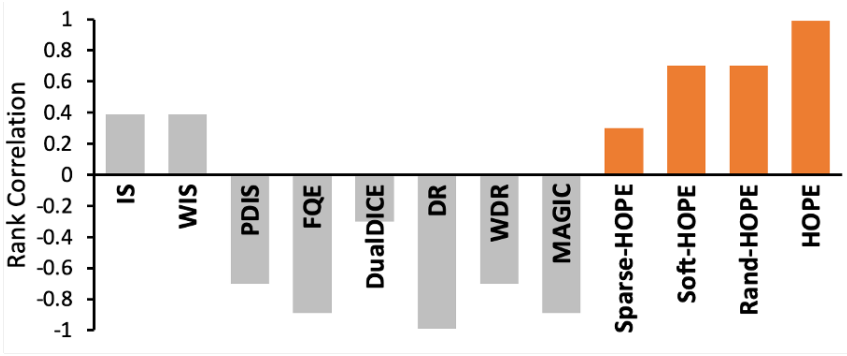


图 3：合成脓毒症环境（γ=0.99）中的平均绝对误差和标准差。IS、加权重要性采样、PHWIS、DualDICE 均无法选出最佳策略。



OPE	Rank Corr.
IS	0.39
WIS	0.39
PDIS	-0.7
FQE	-0.89
DualDICE	-0.3
DR	-0.99
WDR	-0.7
MAGIC	-0.89
Sparse-HOPE	0.3
Soft-HOPE	0.7
Rand-HOPE	0.7
HOPE	**0.99**

图4：现实医疗系统中各策略之间的斯皮尔曼秩相关系数（γ=0.99）。由于轨迹长度差异很大，PHWIS 可能无法产生有意义的结果。

这些指标的定义见附录 B。此外，我们使用以人为中心的显著性检验，来衡量不同策略之间离线策略评估所估计回报的统计显著性。对于每个评估策略，我们按照 [25] 中的介绍，以回合为单位使用自助法。对于电子学习，我们还对诱导策略进行了实证评估。对于许多以人为中心的任务而言，强化学习诱导的策略的一项关键衡量标准，是其是否显著优于当前的专家策略 [81]。因此，我们对通过自助法获得的离线策略评估估计值进行 t检验。该检验用于判断一个策略的离线策略评估估计值均值与另一个策略的均值之间是否存在显著差异。

4.3 脓毒症治疗

脓毒症被定义为机体对感染产生反应而导致的危及生命的器官功能障碍，是死亡率最高的疾病，也是与住院时间相关的成本最为高昂的疾病，在美国每年造成超过240亿美元的成本 [41]。具体而言，脓毒性休克是脓毒症由于循环和/或细胞代谢严重异常而引发的最严重并发症 [7], , 其死亡率最高可达50% [43]。脓毒症治疗是一个极具挑战性的问题，已引发大量研究 [15, 16]。

4.3.1 合成脓毒症环境。 我们使用 Oberst 和 Sontag 的脓毒症模型 [52] 来管理重症监护病房患者的脓毒症。遵循 [49],中的设置，离散观测空间由一个表示糖尿病的二元指标和四项生命体征（心率、血压、氧浓度、血糖水平）组成，这些生命体征的取值属于 {very_high, high, normal, low, very_low} 的子集。仿真环境总共包含 1440 个观测，以及 8 个动作；每个动作通过分别为 {antibiotics, vasopressors, mechanical_ventilation} 中的各项分配二元值（0 或 1）来定义。仿真会持续进行，直到达到最多 T = 5 （时域）个时间步（奖励为 0）、患者死亡（奖励为 -1）或患者出院（奖励为 +1 ）。当所有生命体征均处于正常范围且未接受治疗时，患者出院。当至少三项生命体征超出正常范围时，患者死亡。我们采用行为策略和三个评估策略，遵循 [49]。图 3 显示，

在各评估策略下，HOPE 的平均绝对误差表现最佳。详细结果见附录 B.1。

4.3.2 现实医疗系统。 考虑到治疗患者时的安全性，医疗系统是离线策略评估的另一个重要领域。在我们的实验中，我们使用了从美国一家大型医院收集的电子健康记录（EHRs），涵盖两年内总计221,700次患者就诊。观测空间由15个连续的脓毒症相关临床属性组成，包括七项生命体征
{
H
e
a
r
t
R
at
e,RespiratoryRate,PulseOx,SystolicBP,DiastolicBP,MAP,Temperature}和八项实验室分析物
{Bands,BUN,Lactate,Platelet,Creatinine,BiliRubin,WBC,FIO2}。动作空间的大小为4，包含两种关于{antibiotic_给药、氧气_辅助}的二元治疗选项。临床医生将脓毒症分为四个阶段，并为每个阶段设定奖励：感染（±5）、炎症（±10）、器官衰竭（±20）和脓毒性休克（±50）。当患者进入相应阶段时给予指定的负奖励，而当患者从该阶段恢复时给予正奖励。所收集轨迹的长度范围为1至1160。我们假设临床护理团队经过良好培训并具备医学知识，且遵循脓毒症治疗的标准方案，因此按照 [5]中的介绍学习专家策略。我们使用深度Q网络（DQN） [46]，通过改变超参数来训练策略，并按照附录B.2中的讨论选取其中五个作为评估策略。由于脓毒症研究中的既有工作 [5, 35, 57]将脓毒性休克发生率确定为学习策略的一项重要标准，我们根据离线策略评估给出的估计值，计算各策略排名与其根据脓毒性休克发生率所得实际排名之间的斯皮尔曼秩相关系数。实验细节见附录B.2。

图4展示了 HOPE 和基准方法的结果。灰色阴影列表示基准方法的结果，橙色阴影列表示 HOPE 及其变体的结果。总体而言，HOPE 的秩相关性表现最佳。有趣的是，我们注意到，重要性采样和加权重要性采样优于其他基准方法，尽管先前的理论研究指出，它们可能会受到长时域的影响 [40]。一种可能的原因是，在我们的现实世界设置中，这两种方法从期望误差的降低中获益更多，而由时域带来的方差影响相对较小。在一些长时域环境中也报告了类似发现 [14]。



图5：我们的智能辅导系统图形用户界面。问题陈述窗口（顶部）展示问题内容。对话窗口（右侧中部）显示辅导者向学生发送的消息。学生在响应窗口（右下方）中输入回答，例如编写方程。通过这一过程生成的变量和方程分别显示在变量窗口（左侧中部）和方程窗口（左下方）。

4.4 真实世界智能导师

智能辅导系统（ITS）是模拟人类导师某些方面的计算机系统，并且已被证明能够取得成功 [1, 2, 73]。它们旨在提供教学或反馈，以支持学生的学习，这是强化学习用于提升学生参与度和学习成果的一项重要应用。我们使用一种基于网页的智能辅导系统，为计算机科学专业的学生教授概率知识，涵盖补集定理等十项主要原理。我们收集了美国一所大型公立大学本科计算机科学课程中，七个学期课堂研究期间（包括1,307名学生）的学生-系统交互日志。图5展示了该智能导师的图形用户界面。

在辅导过程中，有许多因素可能决定或反映学生的学习状态，但教育工作者对其中许多因素还不甚了解。因此，为了保持保守性，我们从学生-系统交互日志中提取了可能决定或反映学生学习观测的各种属性。总计提取了142个具有离散值和连续值的属性，可分为以下五组：

(i) **自主性（10个特征）**：学生完成的工作量，例如学生重新开始某个问题的次数；

(ii) **时间情境（29个特征）**：与工作过程相关的时间信息，例如每个步骤的平均用时；

(iii) **问题解决（35个特征）**：有关当前问题解决情境的信息，例如问题难度；

(iv) **表现（57个特征）**：有关学生在问题解决过程中的表现信息，例如正确输入的百分比；

(v) **提示（11个特征）**：有关学生提示使用情况的信息，例如请求提示的总次数。

智能体将做出 10 次决策：对于每道题，智能体将决定学生是应该解答 下一道题、学习导师提供的解答，还是与导师共同 解决问题。所有题目完成后才能获得奖励，其定义为学生的标准化学习增益，分别根据学生在实验前后参加的两次测试分数计算得出 [10], 。共有四种策略被部署到学生使用的 ITS 中，其中包括三种 DQN生成的策略（记为 π_1、π_2、π_3）和一种专家策略（记为 π_{expert}）。使用前六个学期学生的日志数据训练策略，并使用后一个学期的数据进行测试。更多详情见附录 B.3。

图6展示了使用标准验证方法和针对离线策略评估（OPE）的带符号显著性检验，将HOPE（最后一行）与九种原始OPE基准方法（左侧表格顶部区域）进行比较的结果。总体而言，在平均绝对误差（AAE，第2列）、regret@1（第3列）和秩相关（第4列）方面，HOPE（最后一行）的表现优于全部九种基准方法。九种原始OPE基准方法中没有明显的优胜者。表6的第5列展示了各个OPE方法确定的最佳策略。尽管九种原始OPE基准方法均将其他次优策略选为最佳策略，但HOPE是唯一能在实证研究中成功识别 π_1 为最佳策略的方法。更重要的是，九种原始OPE基准方法都预测其选出的最佳策略将显著优于专家策略（第6列），但实证结果（第7列/最后一列）显示，两者之间没有显著差异。然而，使用HOPE进行的离线显著性检验与实证结果完全一致：在OPE t检验（第6列）和实证t检验（第7列）中，π_1 与专家策略均存在显著差异。

4.5 消融实验

为了更好地理解我们所提出的、用于解决现实世界以人为中心的环境中的部分可观测性和即时奖励缺失问题的方法，我们开展了三项消融实验：

(i) Sparse-HOPE。我们所提出方法的一种变体假设初步奖励是稀疏的，并通过最近邻校准即时奖励，称为 Sparse-HOPE。图4和图6表明，就秩相关而言，除重要性采样外，Sparse-HOPE 的表现优于所有基准方法。在现实世界的智能辅导系统中，就平均绝对误差和 regret@1 而言，Sparse-HOPE 的表现优于所有基准方法。这些结果表明，我们所提出的基于最近邻的即时奖励重构方法能够有效估计策略回报。另一方面，Sparse-HOPE 的表现不如 HOPE，这可能说明将奖励视为聚合结果的重要性，以及我们的初步奖励重构方法的有效性。

(ii) Soft-HOPE。我们定义了 HOPE 的另一种变体，称为 Soft-HOPE，该方法假设基于任何观测所做出的决策都可能对最终结果产生同等影响，即

$$V^{\pi}(\text{Soft-HOPE}) = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \sum_{t=1}^{T} \gamma^{t-1} \left(\frac{1}{K} \sum_{k} \hat{r}_{t'}^{(k)}\right). \quad (10)$$

请注意，与公式（6）仅在关键观测上估计基于邻居的即时奖励不同，该方法会对所有观测执行基于邻居的估计。图4显示，就秩相关而言，Soft-HOPE在现实世界医疗系统中的表现优于所有基准方法；就平均绝对误差而言，它在智能辅导中的表现优于所有基准方法。然而，在智能辅导中的regret@1和秩相关方面，其表现不如大多数基准方法。一个可能的原因是使用了平均奖励

	Standard metrics			Human-centric significance test		
OPE	AAE	Regret@1	Rank	π^{Best}	OPE t-test (p)	Empirical t-test (p)
IS	6.42	0.46	**0.8**	π_2	7.071 (0.000)	0.612 (0.543)
WIS	0.22	0.46	0.2	π_2	17.081 (0.000)	0.612 (0.543)
PDIS	6.42	0.46	**0.8**	π_2	7.071 (0.000)	0.612 (0.543)
PHWIS	0.22	0.46	0.2	π_2	17.081 (0.000)	0.612 (0.543)
FQE	0.10	0.46	-0.8	π_2	12.546 (0.000)	0.612 (0.543)
DualDICE	0.33	N/A	N/A	N/A	N/A	N/A
DR	18.5	0.60	**0.8**	π_3	4.893 (0.000)	0.203 (0.840)
WDR	1.63	0.60	-0.8	π_3	8.344 (0.000)	0.203 (0.840)
MAGIC	17.3	0.60	0.6	π_3	20.750 (0.000)	0.203 (0.840)
Sparse-HOPE	0.06	**0.00**	0.4	π_1^*	**10.518 (0.000)**	**2.011 (0.040)**
Soft-HOPE	0.07	**0.00**	**0.8**	π_1^*	1.332 (0.186)	2.011 (0.040)
Rand-HOPE	0.10	**0.00**	**0.8**	π_1^*	**6.902 (0.000)**	**2.011 (0.040)**
HOPE	**0.03**	**0.00**	**0.8**	π_1^*	**11.670 (0.000)**	**2.011 (0.040)**

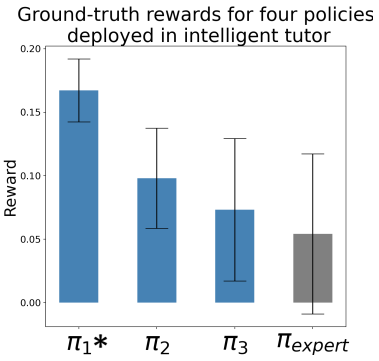


图6：验证离线策略评估的表现（左）以及来自真实世界智能辅导的数据结果（右）（$\gamma = 0.9$）。左：标准指标包括平均绝对误差（AAE）、regret@1和斯皮尔曼秩相关系数（Rank）。每项指标的最佳结果以粗体显示。以人为中心的显著性检验结果包括：由离线策略评估选出的最佳策略、在由离线策略评估选出的最佳策略和专家策略上通过自助法进行的离线显著性检验，以及针对这两个策略的经验显著性检验，显著性水平为 $p < 0.05$。与经验检验一致的离线策略评估显著性检验结果以粗体显示。DualDICE以相同方式估计所有策略的表现，因此无法获得与秩相关的结果。右：四种策略的真实奖励。 π_1 获得了最高的平均奖励，并且是唯一一个与π_{expert} 存在显著差异的策略（$p = 0.04$），如星号所示。

对所有观测进行操作可能会给离线策略评估引入噪声，并削弱其对策略排序的估计。此外，在这两个环境中，它的表现都不如 HOPE，这表明我们定义的关键观测有助于为离线策略评估提取有意义的信息。

（*iii*）*Rand-HOPE*。我们提出的方法的第三种变体是随机选择邻居，而不是使用我们定义的距离来重构即时奖励，我们将其称为 Rand-HOPE。我们重复进行 Rank-HOPE 100 次，并报告平均结果。图4和图6显示，Rand-HOPE 在两个现实环境中均优于所有基准方法。可能的原因是，推断初步即时奖励能够提供比稀疏奖励多得多的有用信息，因此，即使对奖励进行随机平均，其表现也会优于使用稀疏奖励。Rand-HOPE 的表现不如 HOPE，这表明我们定义的距离能够更准确地为电子学习和医疗保健领域的离线策略评估重构即时奖励。

此外，在真实世界智能辅导中，HOPE 及其三种变体是唯一能够成功选出最佳策略的方法。表1进一步展示了 HOPE 相关方法对策略 π_1 和π_{expert} 的估计均值和标准误。HOPE 的估计效果最佳，最接近真实值。

5 结论与社会影响

在这项工作中，我们提出了一种名为 *HOPE* 的方法，用于在具有部分可观测性和聚合奖励的真实世界以人为中心的环境中进行离线策略评估。该方法首先根据历史观测推断初步即时奖励。随后，它使用最近邻方法完整地重建即时奖励。我们还引入了关键观测，这些观测可能影响一条轨迹相较于其他

	π_1 均值$_{\pm se}$	π_{expert} 均值$_{\pm se}$
Sparse-HOPE	$0.094_{\pm 0.01}$	$0.005_{\pm 0.00}$
Soft-HOPE	$0.021_{\pm 0.00}$	$0.004_{\pm 0.00}$
Rand-HOPE	$0.140_{\pm 0.02}$	$-0.023_{\pm 0.02}$
HOPE	$\mathbf{0.176}_{\pm 0.01}$	$\mathbf{0.008}_{\pm 0.00}$
实证结果	$0.167_{\pm 0.02}$	$0.054_{\pm 0.06}$

表1：在真实世界智能辅导中的策略 π_1 和π_{expert} 上使用自助法计算的均值和标准误。HOPE 在两种策略上的估计结果均为最佳。

轨迹的最终结果，以丰富离线策略评估所获得的信息。我们针对离线策略评估中的两项具有挑战性的任务——脓毒症治疗和智能辅导——开展了大量真实世界实验，并采用标准验证方法和以人为中心的显著性检验来验证离线策略评估。结果表明，HOPE 在这两种应用中的表现均优于现有最先进的基准方法。我们的方法有一部分采用了加权重要性采样，这可能会给估计引入方差。我们保持方法简洁，以便轻松分离评估其表现。未来，可以用双重稳健或 DICE 替代加权重要性采样，以降低方差。

所有教育数据和电子健康记录均通过豁免的经机构审查委员会批准的方案以匿名方式获取，并依据既定评分量规进行评分。未收集人口统计数据或课程成绩。所有数据均在机构审查委员会的监督下于研究小组内共享，并经过去标识化处理后自动进行标注。本研究旨在消除因需要更多个性化

干预措施，并为患者制定更稳健的医疗干预。

致谢

本研究得到以下NSF资助项目的支持：用于STEM学习环境个性化教学的集成数据驱动技术（1726550）、CAREER项目：改进交互式学习环境中的自适应决策制定（1651909），以及通过智能导师推广数据驱动技术以改进个性化STEM教学（2013502）。

参考文献

[1] Mark Abdelshiheed、Mehak Maniktala、Song Ju、Ayush Jain、Tiffany Barnes和Min Chi。2021年。为未来学习准备未准备好的学生。载于认知科学学会第*43*届年会论文集，第43卷。[2] Mark Abdelshiheed、Guojing Zhou、Mehak Maniktala、Tiffany Barnes和Min Chi。2020年。元认知与动机：时间意识在为未来学习做准备中的作用。载于认知科学学会第*42*届年会论文集，第42卷。[3] Kenshi Abe和Yusuke Kaneko。2021年。双人零和马尔可夫博弈中的离策略可利用性评估。载于第*20*届国际自主智能体与多智能体系统会议论文集。78–87。[4] Markel Sanz Ausin、Hamoon Azizsoltani、Song Ju、Yeo Jin Kim和Min Chi。2021年。用于延迟强化任务的Infernet：解决时间信用分配问题。载于*2021*年*IEEE*国际大数据会议（*Big Data*）。IEEE，1337–1348。[5] Hamoon Azizsoltani和Yeo Jin。2019年。未观测不等于不存在：使用高斯过程推断跨上下文的即时奖励。载于国际人工智能联合会议第*28*届会议论文集。[6] Andrew Bennett、Nathan Kallus、Lihong Li和Ali Mousavi。2021年。带潜在混杂因素的无限时域强化学习中的离策略评估。载于人工智能与统计国际会议。PMLR，1999–2007。[7] Roger C Bone、Robert A Balk、Frank B Cerra、R Phillip Dellinger、Alan M Fein、William A Knaus、Roland MH Schein和William J Sibbald。1992年。脓毒症和器官衰竭的定义，以及脓毒症创新疗法使用指南。*Chest* 101，6（1992），1644–1655。[8] Anthony R Cassandra。1998年。POMDP应用综述。载于*AAAI 1998*年秋季研讨会“使用部分可观测马尔可夫决策过程进行规划”工作笔记，第1724卷。[9] Yash Chandak、Scott Niekum、Bruno da Silva、Erik Learned-Miller、Emma Brunskill和Philip S Thomas。2021年。通用离策略评估。神经信息处理系统进展 34（2021）。[10] Min Chi、Kurt VanLehn、Diane Litman和Pamela Jordan。2011年。实证评估将强化学习应用于有效且自适应教学策略归纳的效果。用户建模与用户自适应交互 21，1（2011），137–180。[11] Shayan Doroudi、Philip S Thomas和Emma Brunskill。2017年。用于公平策略选择的重要性采样。资助方提交材料（2017）。[12] Mehrdad Farajtabar、Yinlam Chow和Mohammad Ghavamzadeh。2018年。更稳健的双重稳健离策略评估。载于国际机器学习会议。PMLR，1447–1456。[13] Justin Fu、Aviral Kumar、Ofir Nachum、George Tucker和Sergey Levine。2020年。D4RL：深度数据驱动强化学习数据集。*arXiv*预印本 *arXiv:2004.07219*（2020）。[14] Justin Fu、Mohammad Norouzi、Ofir Nachum、George Tucker、Ziyu Wang、Alexander Novikov、Mengjiao Yang、Michael R Zhang、Yutian Chen、Aviral Kumar等。2021年。深度离策略评估基准。*arXiv*预印本 *arXiv:2103.16596*（2021）。[15] Ge Gao、Qitong Gao、Xi Yang、Miroslav Pajic和Min Chi。2022年。强化学习驱动的多变量时间序列分类模式挖掘框架。载于国际人工智能联合会议（*IJCAI*）。[16] Ge Gao、Farzaneh Khoshnevisan和Min Chi。2022年。利用时间感知的就诊内与跨就诊信息重建缺失电子健康记录，以实现脓毒性休克早期预测。载于*2022*年*IEEE*第*10*届医疗保健信息学国际会议（*ICHI*）。IEEE计算机学会，151–162。[17] Ge Gao、Samiha Marwan和Thomas W Price。2021年。利用编程过程数据中的可解释模式进行表现早期预测。载于第*52*届*ACM*计算机科学教育技术研讨会论文集。342–348。[18] Qitong Gao、Ge Gao、Min Chi和Miroslav Pajic。2023年。用于离策略评估的变分潜在分支模型。载于第十一届国际学习表征会议。[19] Qitong Gao、Stephen L Schmidt、Karthik Kamaravelu、Dennis A Turner、Warren M Grill和Miroslav Pajic。2022年。面向基于学习的深度脑刺激控制器。在2022年ACM/IEEE第13届网络物理系统国际会议（ICCPS）论文集中。IEEE，第80–91页。[20] Qitong Gao、Dong Wang、Joshua D Amason、Siyang Yuan、Chenyang Tao、Ricardo Henao、Majda Hadziahmetovic、Lawrence Carin和Miroslav Pajic。2022年。针对不完整观测的梯度重要性学习。国际学习表征会议（2022年）。[21] Omer Gottesman、Joseph Futoma、Yao Liu、Sonali Parbhoo、Leo Celi、Emma Brunskill和Finale Doshi-Velez。2020年。通过突出有影响力的转移，实现强化学习中的可解释离策略评估。载于国际机器学习会议论文集。PMLR，第3658–3667页。[22] Joy Paul Guilford。1950年。心理学与教育学基础统计学。（1950年）。[23] David Hallac、Sagar Vare、Stephen Boyd和Jure Leskovec。2017年。基于托普利茨逆协方差的多变量时间序列数据聚类。载于第23届ACM SIGKDD知识发现与数据挖掘国际会议论文集。第215–223页。[24] Josiah Hanna、Scott Niekum和Peter Stone。2019年。使用估计的行为策略进行重要性采样策略评估。载于国际机器学习会议论文集。PMLR，第2605–2613页。[25] Botao Hao、Xiang Ji、Yaqi Duan、Hao Lu、Csaba Szepesvári和Mengdi Wang。2021年。离策略评估的自助法统计推断。arXiv预印本 arXiv:2102.03607（2021年）。[26] Anna Harutyunyan、Marc G Bellemare、Tom Stepleton和Rémi Munos。2016年。带离策略校正的Q（λ）。载于算法学习理论国际会议论文集。Springer，第305–320页。[27] Daniel G Horvitz和Donovan J Thompson。1952年。有限总体不放回抽样的一种推广。《美国统计学会杂志》47卷260期（1952年），第663–685页。[28] Alexander Irpan、Kanishka Rao、Konstantinos Bousmalis、Chris Harris、Julian Ibarz和Sergey Levine。2019年。通过离策略分类进行离策略评估。《神经信息处理系统进展》32卷（2019年）。[29] Jiming Jiang。2010年。统计学大样本技术。Springer。[30] Nan Jiang和Lihong Li。2016年。强化学习的双重稳健离策略价值评估。载于国际机器学习会议论文集。PMLR，第652–661页。[31] Song Ju、Shitian Shen、Hamoon Azizsoltani、Tiffany Barnes和Min Chi。2019年。通过重要性采样识别经验有效策略及其关键决策。载于EDM（研讨会）。第69–78页。[32] Ramtin Keramati、Omer Gottesman、Leo Anthony Celi、Finale Doshi-Velez和Emma Brunskill。2022年。在离策略策略评估中识别具有相似收益的亚组。载于健康、推断与学习会议论文集。PMLR，第397–410页。[33] Farzaneh Khoshnevisan和Min Chi。2021年。统一领域适应与领域泛化，以提升对少数族裔群体的稳健预测。载于欧洲机器学习与知识发现联合会议论文集。Springer，第521–537页。[34] Yeo-Jin Kim和Min Chi。2018年。时间信念记忆：在RNN训练期间填补缺失数据。载于第27届国际人工智能联合会议（IJCAI-2018）论文集。[35] Matthieu Komorowski、Leo A Celi、Omar Badawi、Anthony C Gordon和A Aldo Faisal。2018年。人工智能临床医生通过学习在重症监护中制定脓毒症最优治疗策略。《自然医学》24卷11期（2018年），第1716–1720页。[36] Solomon Kullback和Richard A Leibler。1951年。关于信息与充分性。《数理统计年刊》22卷1期（1951年），第79–86页。[37] Hoang Le、Cameron Voloshin和Yisong Yue。2019年。约束下的批量策略学习。载于国际机器学习会议论文集。PMLR，第3703–3712页。[38] Timothy P Lillicrap、Jonathan J Hunt、Alexander Pritzel、Nicolas Heess、Tom Erez、Yuval Tassa、David Silver和Daan Wierstra。2015年。通过深度强化学习实现连续控制。arXiv预印本 arXiv:1509.02971（2015年）。[39] Zachary C Lipton、David Kale和Randall Wetzel。2016年。使用RNN直接建模序列中的缺失数据：改进临床时间序列分类。载于医疗保健机器学习会议论文集。PMLR，第253–270页。[40] Qiang Liu、Lihong Li、Ziyang Tang和Dengyong Zhou。2018年。打破时域的束缚：无限时域离策略估计。《神经信息处理系统进展》31卷（2018年）。[41] Vincent Liu、Gabriel J Escobar、John D Greene、Jay Soule、Alan Whippy、Derek C Angus和Theodore J Iwashyna。2014年。来自两个独立队列的脓毒症患者医院死亡情况。《美国医学会杂志》312卷1期（2014年），第90–92页。[42] Travis Mandel、Yun-En Liu、Sergey Levine、Emma Brunskill和Zoran Popovic。2014年。跨表征的离线策略评估及其在教育游戏中的应用。载于AAMAS，第1077卷。[43] G.S Martin、D.M Mannino等。2003年。1979年至2000年美国脓毒症流行病学。《新英格兰医学杂志》（2003年）。[44] Samiha Marwan、Ge Gao、Susan Fisk、Thomas W Price和Tiffany Barnes。2020年。自适应即时反馈可以提高新手的编程参与度以及其继续学习计算机科学的意愿。载于2020年ACM国际计算教育研究会议论文集。第194–203页。

[45] Maja J Mataric。1994年。用于加速学习的奖励函数。载于《机器学习论文集 1994》。Elsevier，第181–189页。[46] Volodymyr Mnih、Koray Kavukcuoglu、David Silver、Andrei A Rusu、Joel Veness、Marc G Bellemare、Alex Graves、Martin Riedmiller、Andreas K Fidjeland、Georg Ostrovski等。2015年。通过深度强化学习实现人类水平控制。《自然》518卷 7540期（2015年），第529–533页。[47] Rémi Munos、Tom Stepleton、Anna Harutyunyan和Marc Bellemare。2016年。安全高效的离策略强化学习。《神经信息处理系统进展》29卷（2016年）。[48] Ofir Nachum、Yinlam Chow、Bo Dai和Lihong Li。2019年。DualDICE：与行为策略无关的折扣平稳分布校正估计。《神经信息处理系统进展》32卷（2019年）。[49] Hongseok Namkoong、Ramtin Keramati、Steve Yadlowsky和Emma Brunskill。2020年。未观测混杂因素下序贯决策的离策略策略评估。《神经信息处理系统进展》33卷（2020年），第18819–18831页。[50] Allen Nie、Yannis Flet-Berliac、Deon Richmond Jordan、William Steenbergen和Emma Brunskill。2022年。有限数据条件下离线强化学习的数据高效管线。载于《神经信息处理系统进展》。[51] Xinkun Nie、Emma Brunskill和Stefan Wager。2021年。学习何时治疗策略。《美国统计学会杂志》116卷533期（2021年），第392–409页。[52] Michael Oberst和David Sontag。2019年。使用Gumbel-max结构因果模型进行反事实离策略评估。载于国际机器学习会议论文集。PMLR，第4881–4890页。[53] Cosmin Paduraru。2013年。马尔可夫决策过程中的离策略评估。（2013年）。[54] Emily Miller Payne、Russ Hodges和Elda Patricia Hernandez。2017年。21世纪学习中心的人口结构变化与需求评估。《学习支持评论》22卷1期（2017年），第21–36页。[55] Michael JD Powell和J Swann。1966年。加权均匀采样：一种降低方差的蒙特卡罗技术。《IMA应用数学杂志》2卷3期（1966年），第228–236页。[56] Doina Precup。2000年。用于离策略策略评估的资格迹。《计算机科学系教师出版物系列》（2000年），第80页。[57] Aniruddh Raghu、Matthieu Komorowski、Imran Ahmed、Leo Celi、Peter Szolovits和Marzyeh Ghassemi。2017年。用于脓毒症治疗的深度强化学习。arXiv预印本 arXiv:1711.09602（2017年）。[58] Desik Rengarajan、Gargi Vaidya、Akshay Sarvesh、Dileep Kalathil和Srinivas Shakkottai。2022年。利用离线示范指导稀疏奖励强化学习。arXiv预印本 arXiv:2202.04628（2022年）。[59] Pranab K Sen和Julio M Singer。1994年。统计学大样本方法：应用导论。第25卷。CRC出版社。[60] Shitian Shen和Min Chi。2016年。强化学习：越早越好，还是越晚越好？载于2016年用户建模、自适应与个性化会议论文集。第37–44页。[61] Chengchun Shi、Masatoshi Uehara、Jiawei Huang和Nan Jiang。2022年。混杂部分可观测马尔可夫决策过程离策略评估的极小极大学习方法。载于国际机器学习会议论文集。PMLR，第20057–20094页。[62] Mervyn Singer、Clifford S Deutschman、Christopher Warren Seymour、Manu Shankar-Hari、Djillali Annane、Michael Bauer、Rinaldo Bellomo、Gordon R Bernard、Jean-Daniel Chiche、Craig M Coopersmith等。2016年。脓毒症和脓毒性休克的第三次国际共识定义（Sepsis-3）。《美国医学会杂志》315卷8期（2016年），第801–810页。[63] Charles Spearman。1987年。两事物之间关联的证明与测量。《美国心理学杂志》100卷3/4期（1987年），第441–471页。[64] Guy Tennenholtz、Uri Shalit和Shie Mannor。2020年。部分可观测环境中的离策略评估。载于AAAI人工智能会议论文集，第34卷。第10276–10283页。[65] Joel Tetreault和Diane Litman。2006年。使用强化学习比较口语对话中状态特征的效用。载于NAACL人类语言技术会议主会场论文集。第272–279页。[66] Philip Thomas和Emma Brunskill。2016年。强化学习的数据高效离策略策略评估。载于国际机器学习会议论文集。PMLR，第2139–2148页。[67] Philip S Thomas。2015年。安全强化学习。（2015年）。[68] Philip S Thomas、Scott Niekum、Georgios Theocharous和George Konidaris。2015年。使用Ω回报进行策略评估。《神经信息处理系统进展》28卷（2015年）。[69] Emanuel Todorov、Tom Erez和Yuval Tassa。2012年。MuJoCo：用于基于模型控制的物理引擎。载于2012年IEEE/RSJ智能机器人与系统国际会议论文集。IEEE，第5026–5033页。[70] Lisa Torrey和Matthew Taylor。2013年。预算有限条件下的教学：强化学习中为智能体提供建议的智能体。载于2013年自主智能体与多智能体系统国际会议论文集。第1053–1060页。[71] Masatoshi Uehara、Jiawei Huang和Nan Jiang。2020年。用于离策略评估的极小极大权重与Q函数学习。载于国际机器学习会议论文集。PMLR，第9659–9668页。[72] Masatoshi Uehara、Haruka Kiyohara、Andrew Bennett、Victor Chernozhukov、Nan Jiang、Nathan Kallus、Chengchun Shi 和 Wen Sun。2022年。《部分可观测马尔可夫决策过程中的未来依赖的基于价值的离策略评估》。*arXiv*预印本 *arXiv:2207.13081*（2022年）。[73] Kurt VanLehn。2006年。《辅导系统的行为》。《国际教育人工智能杂志》16，3（2006年），227–265。[74] Cameron Voloshin、Hoang M Le、Nan Jiang 和 Yisong Yue。2019年。《强化学习离策略策略评估的实证研究》。*arXiv*预印本 *arXiv:1911.06854*（2019年）。[75] Tengyang Xie、Yifei Ma 和 Yu-Xiang Wang。2019 年。《利用边际重要性采样实现强化学习的最优离策略评估》。神经信息处理系统进展32（2019年）。[76] Mengjiao Yang、Bo Dai、Ofir Nachum、George Tucker 和 Dale Schuurmans。2022年。《不确定性下的离线策略选择》。载于人工智能与统计国际会议。PMLR，4376–4396。[77] Xi Yang、Yuan Zhang 和 Min Chi。2021年。《用于疾病进展建模的多序列时间感知序列划分》。载于*IJCAI*。[78] Xi Yang、Guojing Zhou、Michelle Taub、Roger Azevedo 和 Min Chi。2020年。《通过EM逆强化学习进行学生子类型划分》。国际教育数据挖掘学会（2020年）。[79] Changchang Yin、Ruoqi Liu、Dongdong Zhang 和 Ping Zhang。2020年。《通过时间感知多模态自编码器识别脓毒症亚表型》。载于第*26*届*ACM SIGKDD*国际会议论文集：知识发现与数据挖掘。862–872。[80] Guojing Zhou、Hamoon Azizsoltani、Markel Sanz Ausin、Tiffany Barnes 和 Min Chi。2022年。《利用粒度：用于教学策略归纳的分层强化学习》。《国际教育人工智能杂志》32，2（2022年），454–500。[81] Guojing Zhou、Xi Yang、Hamoon Azizsoltani、Tiffany Barnes 和 Min Chi。2020年。《通过数据驱动的解释改进学生-系统交互：由分层强化学习诱导的教学策略》。载于第*28*届*ACM*用户建模、自适应与个性化会议论文集。284–292。

算法 2： HOPE算法。

输入： 历史数据 D ～ β。目标策略 π。参数 h 和 K.

开始：

1：根据（1）推断 $\tilde{r}$。
2：初始化一个空集 O^*。
3：训练 $Q^{\beta}(o, a)$，并获得关键观测集 O^* 给定 h。
4：**对于**每个 τ（i）∈ D **执行**
5：　**对于**每个 o（i）$_t$ ∈ τ（i）**执行**
6：　　**if** $o_t^{(i)} \in O^*$ **则**
7：　　按照算法1查找 K 个 o（i）$_t$ 最近邻。
8：　　计算平均即时奖励 $\bar{r}(i)\,t$ 依据 4。
9：　　**结束 if**
10：　　重构即时奖励 $\hat{r}(i)\,t$ 依据 2。
11：　**结束 for**
12：**结束 for**

一个 HOPE

下文中，K 最近邻的距离度量既适用于离散状态空间，也适用于连续状态空间。为提高泛化能力，可以使用一些通用距离度量，例如 Kullback–Leibler（KL）散度 [36], 来计算 $similarity(o(i)\,o(k))$和 $similarity(a(i)\,a(k))$。对于离散状态空间 O ={o_0，……，o_M}，

$$similarity(o^{(i)}, o^{(k)}) = \sum_{o_m \in s^{(i)}} \Phi^{(k)}(o_m) \log(\frac{\Phi^{(k)}(o_m)}{\Phi^{(i)}(o_m)}), \qquad (11)$$

其中，$\Phi(k)(o_m)$ 和 $\Phi(i)(o_m)$ 分别是关于 $\tau(k)$ 和 $\tau(i)$ 上观测 o_m概率分布。同样，$similarity(a(i)\,a(k))$ $\sum_{a_l \in a^{(i)}} \Phi_{(k)}(a_l) \log(\frac{\Phi^{(k)}(a_l)}{\Phi^{(i)}(a_l)})$，对于动作空间 A ={$a_1, ..., a_L$}。

对于连续观测空间，我们通过刻画观测各特征之间的依赖关系来对其进行离散化。具体而言，我们将与其时间信息（即从轨迹开始经过的时间）相关联的观测分组为M 个簇，其中每个观测都会从簇集合中被分配一个离散值。我们选择将目标表述为基于托普利茨逆协方差的聚类问题 [23] 中定义的形式，并采用其针对多序列时间感知轨迹 [77], 的实现，因为通过考虑图连通性可以以基于模型的方式求解该问题。

B 实验细节

标准验证指标

绝对误差 策略的绝对误差定义为实际值与估计值之差：

$$AE = |V^{\pi} - \hat{V}^{\pi}| \qquad (12)$$

ˆ 其中 V^{π}表示策略 π的实际值，而 V^{π}表示 π的估计值。

regret@1 regret@1 是实际最佳策略的价值与根据估计值选出的最佳策略的实际价值之间的（归一化）差值

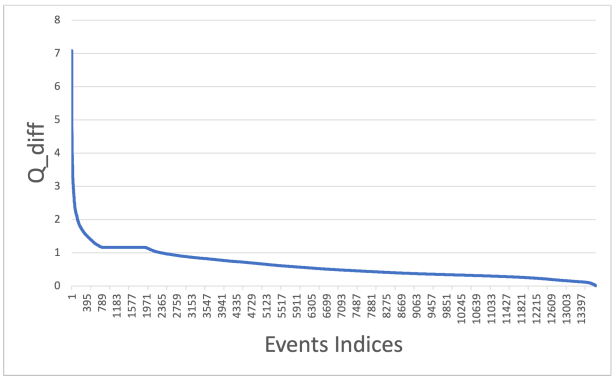


图 7：智能导师中 Q^{π}的最大差异。我们使用肘部法并选择 h = 1.1。

。它可以定义为：

$$R1 = (\max_{i \in 1:P} V_i^{\pi} - \max_{j \in \text{best}(1:P)} V_j^{\pi}) / \max_{i \in 1:P} V_i^{\pi} \qquad (13)$$

其中，best(1 : P) 表示根据估计值 ˆV^{π} 衡量时，在 P个策略集合中最佳策略的索引。

秩相关 秩相关用于衡量各策略的估计值与实际值的序数排名之间的斯皮尔曼秩相关系数：

$$\rho = \frac{Cov(\text{rank}(V_{1:P}^{\pi}), \text{rank}(\hat{V}_{1:P}^{\pi}))}{\sigma(\text{rank}(V_{1:P}^{\pi}))\sigma(\text{rank}(\hat{V}_{1:P}^{\pi}))} \qquad (14)$$

其中，rank($V^{\pi}_{1:P}$)表示各策略实际值ˆ的序数排名，而rank($V^{\pi}_{1:P}$)表示各策略估计值的序数排名。

实验设置我们使用 Python 实现了所提出的方法。所有实验均在一台配备四块 NVIDIA TITAN Xp / 12GB RAM、两颗六核 Xeon E5-2620 @2 GHz CPU 的 CentOS7.8（64 位）机器上运行。

参数设置。 在本研究中，我们使用 K = 5，遵循相关的基于邻居的方法 [21], ，以确保本研究结果具有普适性。对于用于识别关键观测的 Q，我们使用 DQN 学习 Q值，其中包含三个隐藏层，每层有 256 个节点。隐藏层采用 ReLU激活，并应用了 20% 的丢弃率。输出层采用线性激活。对于每个观测，都会计算 $Q(o, a)$ a ∈ A 的最大差值。在仿真中，由于时域相较于现实轨迹较短，我们假设每个观测都可以是关键观测。在现实环境中，阈值通过肘部法选取。图 7 展示了在我们的真实世界智能辅导中选择 h 的示例。

详细基准设置 对于 FQE，我们采用了不同的收敛 ϵ ={$1e-3$、$1e-4$、$1e-5$、$4e-3$}，遵循 [74] 并报告平均绝对误差最小（或秩相关最高）的结果。对于 DualDICE，我们使用其原论文中的开源代码，采用不同的随机种子重复运行三次，并报告平均绝对误差最小的结果（或在现实医疗系统中秩相关最高的结果）。

B.1 ICU患者模拟

模拟器模拟了 10000 条轨迹。对于每种离线策略评估（OPE），它都按照 [49] 估计了每个策略的表现，包括{optimal_策略、使用_抗生素的策略、不使用_抗生素的策略}。在此，我们使用自助法对评估策略进行 500 次离线策略评估估计，并计算估计值的平均值及标准差。

B.2 现实医疗系统

B.2.1 数据预处理。 **标签**。 医院提供了为期两年的电子健康记录，其中包括 221,700 次就诊记录，涵盖 35 个静态变量（如性别、年龄和既往病史）以及 43 个时间变量，包括生命体征、实验室分析物和治疗措施。我们的研究人群为疑似感染患者；疑似感染的判定依据是使用了任何类型的抗生素、抗病毒药物、抗菌药物、抗寄生虫药物或抗真菌药物，或 PCR（即时快速检测）结果呈阳性。根据《脓毒症与脓毒性休克第三次国际共识定义》 [62],，我们的医学专家将符合以下任一条件的情况认定为脓毒性休克：

- 持续性低血压表现为两次连续测量结果均符合条件（间隔 ≤ 30 分钟）。- 收缩压（SBP） < 90 mmHg - 平均动脉压（MAP） < 65 mmHg - 在 8 小时内收缩压下降 ≥ 40 mmHg • 任何血管加压药给药。

根据专家脓毒症诊断规则与国际疾病分类第九版（ICD-9）的交集，从电子健康记录中识别出3,499次脓毒性休克阳性就诊和81,398次阴性就诊；专家规则与ICD-9标签不一致的36,122次就诊在本研究中被排除。通过排除入院时即患有脓毒性休克的就诊和长期住院就诊，并随后从非休克就诊中进行分层随机抽样，获得了2,205次休克就诊，同时保持年龄、性别、种族和住院时长的分布一致。最终数据由4,410次就诊构成，其中休克与非休克就诊的比例相等。

观测。 为近似患者观测，根据脓毒症诊断规则选取了15个脓毒症相关属性。在我们的数据中，15个脓毒症相关属性的平均缺失率为78.6%。我们没有删除稀疏属性，也没有按照固定时间间隔进行重采样，因为医学专家建议的属性对于脓毒症治疗决策至关重要，而且电子健康记录的时间缺失模式也提供了患者观测的信息。缺失值通过时间信念记忆 [34] 结合缺失指示变量 [39]进行插补。

动作。 对于动作，我们考虑了两种医疗治疗措施：抗生素给药和氧疗。需要注意的是，这两种治疗措施可以同时实施，因此总共有四种动作。通常，根据粒度的不同，治疗措施会分别采用离散动作空间和连续动作空间进行表示。例如，判断是否施用某种药物属于离散决策，而药物剂量则是连续的。连续动作空间主要通过基于策略的强化学习模型（如演员-评论家模型） [38],进行处理，并且通常仅适用于在线强化学习。由于我们无法在与实际患者在线交互时搜索连续动作空间，因此我们专注于离散动作。

在与实际患者交互时，我们无法搜索连续动作空间，因此我们专注于离散动作。此外，在本研究中，强化学习智能体旨在告知医生何时以及应对患者采取何种治疗措施，而不是建议药物的最佳剂量或氧气控制的持续时间，后者需要更复杂的考量。

奖励。 两位在脓毒症领域均拥有超过20年经验的资深临床医生，根据脓毒症阶段的严重程度指导定义了奖励函数。奖励定义如下：感染 [-5], 炎症 [-10], 器官衰竭 [-20], 以及脓毒性休克 [-50]。每当患者从其中任一阶段恢复时，便重新获得该阶段对应的正奖励。

数据被划分为80%（根据患者就诊记录中首次事件发生时间排序后较早的80%）用于训练，以及（较晚的）20%用于测试，因为离线策略评估最常见的任务是使用历史数据验证策略，然后将选定的策略用于测试。

策略 我们按照 [14, 24]中的方法，使用行为克隆估计行为策略。评估策略采用离策略DQN算法进行训练，并使用不同的学习率。

B.2.2 脓毒性休克发生率。 由于强化学习智能体无法直接与患者互动，因此策略诱导和评估都只能依赖离线数据。与先前的研究 [5, 35, 57], 类似，诱导出的策略使用脓毒性休克发生率进行评估。其背后的假设 [57]是：当脓毒性休克预防策略确实有效时，患者轨迹中的实际治疗与诱导策略越一致，患者发生脓毒性休克的可能性就越低；反之，患者轨迹中的实际治疗与诱导策略越不一致（差异越大），患者发生脓毒性休克的可能性就越高。具体而言，我们测量了与智能体策略的一致率，即 $a \in [0, 1]$ 在一次就诊中，与智能体策略一致的事件数占事件总数的比例； $a = 0$ 如果一次就诊轨迹中的实际治疗措施与智能体建议完全不同，以及 $a = 1$ 如果二者相同。根据一致率，计算平均脓毒性休克发生率，即具有相应一致率的就诊中发生休克的就诊次数所占的比例 $\geq a$。如果智能体策略确实有效，那么实际执行的治疗措施与智能体策略越一致，患者发生脓毒性休克的可能性就越低。该指标首次用于 [57]。

B.3 真实世界智能导师

B.3.1 数据收集与预处理。 我们的数据包含在七个学期的课堂研究中收集的、共计1,307名学生与基于网页的智能辅导系统的交互日志。在研究期间，所有学生使用同一个智能辅导系统，遵循相同的一般流程，学习相同的训练材料，并完成相同的训练题。所有学生都经历了相同的四个阶段：1）阅读教科书；2）前测；3）使用智能辅导系统进行学习；4）后测。在阅读教科书阶段，学生阅读每个原理的一般性描述，复习示例，并解决一些训练题，以熟悉智能辅导系统。随后，学生进行前测，其中共包含14道单原理和多原理问题。学生不会收到关于其答案的反馈，也不允许返回之前的问题（后测同样如此）。接下来，学生使用智能辅导系统，在其中他们会收到

相同的10道题，且顺序相同。之后，学生进行包含20道题的后测，其中14道题与前测中的题目同构，其余题目为非同构的多原理问题。测试按照相同的评分标准进行自动评分。测试分数被归一化到 $[0, 1]$的范围内。**奖励**。 没有即时奖励，只有经验评估矩阵（即延迟奖励），该矩阵即学生的标准化学习增益（NLG）。NLG衡量学生的学习增益，而不受其初始能力的影响。NLG定义为： $NLG = \frac{score_{posttest} - score_{pretest}}{\sqrt{1 - score_{pretest}}}$，其中1表示最大值

分别是在使用智能辅导系统之前和之后进行的前测和后测成绩。

策略。 这项研究历时七个学期，前六个学期的数据是在专家策略下收集的，第七个学期的数据则是在四种不同策略下收集的（三种策略是强化学习诱导策略，一种是专家策略）。专家策略会随机选择动作。三种强化学习诱导策略采用具有不同学习率的离策略DQN算法进行训练。